Jednoduchá lineární regrese II

Opakování

Studujeme regresní model

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, kde

$\mathbf{y} = (y_1, \ldots, y_n)'$ - vektor pozorování závisle proměnné veličiny Y,

$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \ldots & f_p(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ 1 & f_1(x_n) & \ldots & f_p(x_n) \end{pmatrix}$ - regresní matice

(předpokládáme, že $h(\mathbf{X}) = p+1 < n$)

$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ - vektor regresních parametrů,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ - vektor náhodných odchylek, pro který platí $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

$\mathbf{X'X\beta} = \mathbf{X'y}$ - systém normálních rovnic

$\mathbf{b} = (\mathbf{X'X})^{-1} \mathbf{X'y}$ – odhad vektoru $\boldsymbol{\beta}$ získaný metodou nejmenších čtverců

$\hat{\mathbf{y}} = \mathbf{Xb}$ – vektor regresních odhadů (vektor predikce)

$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ - vektor reziduí

Vlastnosti odhadu **b**:

- odhad **b** je lineární, neboť je vytvořen lineární kombinací pozorování $y_1, \ldots, y_n$ s maticí vah $(\mathbf{X'X})^{-1}\mathbf{X'}$;
- odhad **b** je nestranný, tj. $E(\mathbf{b}) = \boldsymbol{\beta}$;
- odhad **b** má varianční matici $\text{var}\,\mathbf{b} = \sigma^2(\mathbf{X'X})^{-1}$;
- odhad $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X'X})^{-1})$;
- odhad **b** je nejlepší nestranný lineární odhad vektoru $\boldsymbol{\beta}$.

Součty čtverců a index determinace:

$S_E = \mathbf{e}'\mathbf{e}$ … reziduální součet čtverců (podíl $s^2 = \frac{S_E}{\quad}$ je odhad rozptylu σ^2)

$S_R = (\hat{\mathbf{y}}-\mathbf{m}_2)'(\hat{\mathbf{y}}-\mathbf{m}_2)$ … regresní součet čtverců, kde $\mathbf{m}_2$ je sloupcový vektor průměrů závisle proměnné veličiny Y

$S_T = (\mathbf{y}-\mathbf{m}_2)'(\mathbf{y}-\mathbf{m}_2)$ … celkový součet čtverců

Platí $S_T = \quad +$

$ID = \quad =$ - index determinace ($0 <$ $<$), udává, jakou část variability Y lze vysvětlit zvolenou regresní funkcí

Intervaly spolehlivosti pro regresní parametry

100(1- α)% interval spolehlivosti pro β_j má meze: $b_j \pm$ $\quad b_j$,

kde $s_{b_j} =$ $\sqrt{v_{jj}}$ je směrodatná chyba odhadu b_j, v_{jj} je j-tý diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$, j = 0, 1, …, p

Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti α testujeme H_0: $\beta_1 = \ldots = \beta_p = 0$ proti H_1: $\beta_1 \neq \ldots \neq \beta_p \neq 0$.

Testová statistika: $F = \frac{S_R/p}{\quad}$ má rozložení F(p, n-p-1), pokud H_0 platí. Kritický obor: $W =$ $\quad$.

$F \in W$ H_0 zamítáme na hladině významnosti α.

Testování významnosti regresních parametrů (dílčí t-testy)

Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu

H_0: $\beta_j = 0$ proti H_1: $\beta_j \neq 0$.

Testová statistika: $T_j = \frac{b_j}{\quad}$ má rozložení t(n-p-1), pokud H_0 platí. Kritický obor: $W =$ $\quad$.

$t_j \in W$ H_0 zamítáme na hladině významnosti α.

Nové poznatky

Interval spolehlivosti pro teoretickou regresní funkci

Nechť x_0 je pevně zvolená hodnota nezávisle proměnné veličiny X.

Vytvoříme vektor $\mathbf{x}_0 = (f_1(x_0), \ldots, f_p(x_0))'$ a zabýváme se lineární kombinací $\mathbf{x}_0'\boldsymbol{\beta}$ složek vektoru regresních parametrů, tj.

hodnotou $m(x_0; \beta_0, \beta_1, \ldots, \beta_p) = \sum_{j=0}^{p} \beta_j f_j(x_0)$ teoretické regresní funkce v bodě x_0.

100(1- α)% interval spolehlivosti pro $\mathbf{x}_0'\boldsymbol{\beta}$, tj. pro hodnotu regresní funkce $m(x_0; \beta_0, \beta_1, \ldots, \beta_p)$ . . . . . .

$$\mathbf{x}_0'\mathbf{b} \pm t_{1-\alpha/2}(n-p)\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

Při spojitě se měnícím x_0 vytvoří meze tohoto intervalu spolehlivosti tzv. pás spolehlivosti kolem regresní funkce.
Tento pás spolehlivosti však nelze interpretovat tak, že pokrývá celou regresní funkci s pravděpodobností 1- α . . . . .
ukazuje na šířku intervalu spolehlivosti pro vypočtenou hodnotu z modelu pro zvolenou hodnotu argumentu x_0.

Příklad: U automobilu Škoda 120 byla změřena spotřeba benzínu (v l/100 km) v závislosti na rychlosti (v km/h).

rychlost X	40	50	60	70	80	90	100	110
spotřeba Y	5,7	5,4	5,2	5,2	5,8	6,0	7,5	8,1

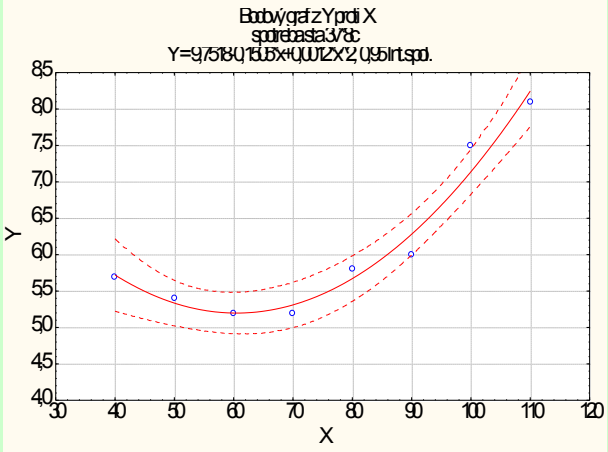
Vhodným modelem je regresní parabola $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$. Odhadněte její parametry a najděte 95% pás spolehlivosti kolem regresní funkce.

Řešení:

N=8	Výsledky regrese se závislou proměnnou R= ,98403165 R2= ,96831829 Upravené F(2,5)=76,410 p<,00018 Směrod. chyba					
	b*	Sm.chy z b*	b	Sm.chy z b	t(5)	p-hod
Abs.čl			9,751	0,945	10,31	0,000
X	-3,38	0,602	-0,150	0,026	-5,61	0,002
Xkv	4,22	0,602	0,001	0,000	7,01	0,000

Spotřeba = 9,751786 – 0,150536*rychlost + 0,001244*rychlost^2

Získání 95% pásu spolehlivosti kolem regresní funkce: Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zvolíme Proložení Polynomiální (implicitně je nastaveno na polynom 2. stupně, lze měnit na záložce Možnosti 2) – zapneme Regresní pásy Spolehl. – OK.



Predikční interval spolehlivosti

V případě, kdy chceme zkonstruovat 100(1- α)% interval spolehlivosti nikoli pro hodnotu regresní funkce, ale pro i-tou predikovanou hodnotu $\ddot{y}_i$ (tzv. predikční interval), dostaneme meze

$$\mathbf{x}_0'\mathbf{b} \pm \dots$$

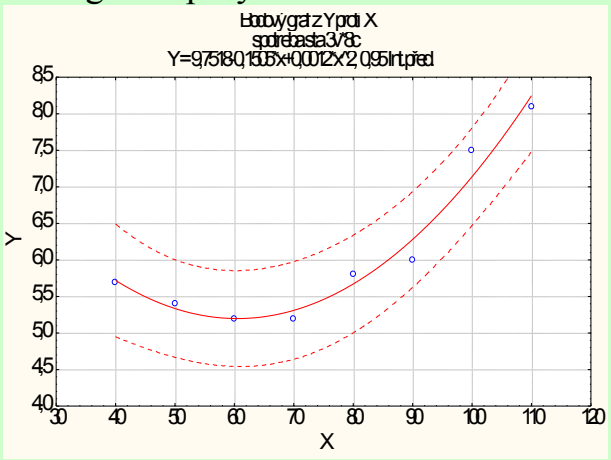
Vidíme, že tento predikční interval je širší než předešlý interval spolehlivosti.

Je to interval, který nás informuje o tom, v jakém rozsahu můžeme očekávat jedno další pozorování s pravděpodobností aspoň 1- α

Při spojitě se měnícím $\mathbf{x}_0$ vytvoří meze tohoto predikčního intervalu spolehlivosti tzv. predikční pás spolehlivosti kolem regresní funkce.

Příklad: Pro model regresní paraboly z předešlého příkladu sestrojte 95% predikční pás spolehlivosti kolem regresní funkce.

Řešení: Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zvolíme Proložení Polynomiální (implicitně je nastaveno na polynom 2. stupně) – zapneme Regresní pásy Predikce – OK.



Chceme-li mít v jednom obrázku zakresleny oba typy pásů, postupujeme takto: ve vytvořeném grafu 2x klikneme na pozadí – vybereme Regresní pásy – Přidat nový pár pásů – OK.

Test adekvátnosti regresního modelu

Nechť hodnoty závisle proměnné veličiny Y jsou roztříděny do $r \geq 3$ skupin podle variant $x_{[1]}, ..., x_{[r]}$ nezávisle proměnné veličiny X. Označme n_i počet pozorování v i-té skupině, i = 1, …, r, přičemž aspoň jedna skupina má více než jedno pozorování. Budeme předpokládat, že každá skupina hodnot má normální rozložení a že všechny skupiny mají týž rozptyl. Všech pozorování je n. Průměr hodnot v i-té skupině označme M_i a průměr všech hodnot označme M.
Charakter závislosti Y na X popíšeme regresní funkcí $m(x;\beta_0,\ldots,\beta_p)$ a budeme se zabývat testováním hypotézy, zda je tato regresní funkce vhodným modelem pro naše data. Při testování budeme potřebovat tyto součty čtverců:

celkový součet čtverců $S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - M)^2$,

skupinový součet čtverců $S_A = \sum_{i=1}^{r} n_i (M_i - M)^2$,

regresní součet čtverců $S_R = \sum_{i=1}^{r} n_i (\hat{y}_i - M_i)^2$.

Testová statistika: $F = \frac{S_A/(r-p-1)}{S_R/(n-r)}$ se řídí rozložením F(r-p-1,n-r), jestliže H_0 platí.

Kritický obor: $W = \langle F_{1-\alpha}(r-p-1, n-r), \infty)$
$F \in W$ => na hladině významnosti α zamítáme hypotézu, že funkce $m(x;\beta_0,\ldots,\beta_p)$ je vhodným regresním modelem závislosti Y na X.

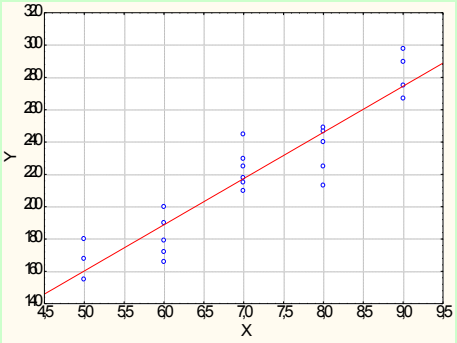
Těsnost závislosti Y na X vyjádřenou skupinovými průměry měří poměr determinace $P^2 = S_A/S_T$.
Nabývá hodnot z intervalu <0,1>. Čím je poměr determinace bližší jedné, tím je závislost silnější, čím je bližší nule, tím je závislost slabší.

Příklad: Máme k dispozici údaje o cenách 23 náhodně vybraných domů (veličina Y - v tisících $) a počtu jejich pokojů (veličina X) v jednom americkém městě.

počet pokojů	cena
5	155,168,180
6	166,172,179,190,200
7	210,215,218,225,230,245
8	213,225,240,247,249
9	267,275,290,298

Závislost ceny domu na počtu pokojů popište regresní přímkou.
Na hladině významnosti 0,05 testujte hypotézu, že přímka je vhodným regresním modelem pro tato data.
Těsnost závislosti vyjádřete poměrem determinace.
Znázorněte data s proloženou regresní přímkou.

Řešení: Empirická regresní přímka má tvar $y = 17{,}2885 + 28{,}5851\ x$,
$S_R = 30907{,}9041$, $S_T = 35870{,}6087$, $S_A = 32474{,}1087$,
$F = \frac{32474{,}1087}{30907{,}9041/18} = 18{,}876$, $F_{0,95}(3,18) = 3{,}161$,
kritický obor $W = \langle 3{,}161, \infty)$. Jelikož $F \in W$, nezamítáme na hladině významnosti 0,05 hypotézu, že přímka je vhodným regresním modelem.
Poměr determinace: $P^2 = S_A/S_T = 32474{,}1087/35870{,}6087 = 0{,}9053$,
tedy závislost ceny domu na počtu pokojů je v daném datovém souboru značně silná.



Řešení v systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 23 případy:

	1 X	2 Y
1	5	15
2	5	16
3	5	18
4	6	16
5	6	17
6	6	17
7	6	19
8	6	20
9	7	21
10	7	21
11	7	21
12	7	22
13	7	23
14	7	24
15	8	21
16	8	22
17	8	24
18	8	24
19	8	24
20	9	26
21	9	27
22	9	29
23	9	29

Odhadneme parametry regresní přímky:

	Výsledky regrese se závislou proměnnou R= ,92825096 R2= ,86164984 Upravené F(1,21)=130,79 p<,00000 Směrod. chyba					
N=23	Beta	Sm.chy beta	B	Sm.chy B	t(21)	Úroveň
Abs.čl			17,28	18,00	0,96(	0,347
X	0,928	0,081	28,58	2,49	11,43	0,000

Cena = 17,28851 + 28,5806*počet pokojů

Sestavíme tabulku ANOVA:
Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

	Analýza rozptylu (ceny_bytu				
Efekt	Souč čtverc	sv	Prům čtverc	F	Úroveň
Regre	3090	1	3090	130,7	0,000
Rezid	4962	2	236,		
Celk.	3587(				

Vidíme, že $S_R = 30907{,}9$, $S_T = 35870{,}61$

Provedeme jednofaktorovou analýzu rozptylu, abychom získali skupinový součet čtverců:
Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – Y, Grupovací - X – OK – OK – Analýza rozptylu.

	Analýza rozptylu (ceny_bytu.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměn	SC efekt	SV efekt	PC efekt	SC chyba	SV chyb	PC chyba	F	p
Y	32472	4	8118,	3396,	1	188,6	43,02	0,000

Zde najdeme $S_A = 32474{,}11$.

Vypočteme testovou statistiku $F = \frac{S_A / r}{S_T / n} = \frac{32474{,}11 / 4}{35870{,}61 / 184} = 41{,}76$

a najdeme kritický obor $W = \langle 3{,}161, \infty)$. Jelikož $F \in W$, nezamítáme na hladině významnosti 0,05 hypotézu, že přímka je vhodným regresním modelem.

Test adekvátnosti modelu pomocí Obecných regresních modelů

Zadáme data a použijeme cestu:
Statistiky – Pokročilé lineární/nelineární modely – Obecné regresní modely – Jednorozměrná regrese - OK – na záložce Možnosti zaškrtneme Kvalita proložení – OK – Závislá Y, Spoj. nezáv. prom. X – OK – Více výsledků – Celkové R – ve stromové struktuře vlevo vybereme Test kvality modelu.

Závislá Proměn	Test kvality modelu (ceny_bytu.sta)										
	SČ Rezid	sv Rezid	PČ Rezid	SČ Chyba	sv Chyb	PČ Chyba	SČ Kv prolož	SV Kv prolož	PČ Kv prolož	ta F	p
Y	4962,	2	236,3	3396,	1	188,6	1566,	3	522,0	2,766	0,071

Čitatel testové statistiky F je roven 1566,205 a je uveden ve sloupci Kvalita proložení.
Jmenovatel testové statistiky F je roven 3396,5 a je uveden ve sloupci SČ Chyba.
Hodnota testové statistiky je 2,767 a odpovídající p-hodnota je 0,0717. Na hladině významnosti 0,05 tedy nemůžeme zamítnout hypotézu, že přímka je vhodným modelem k popisu závislosti ceny domu na počtu pokojů.

Regresní přímka a její vlastnosti

Uvažujeme regresní model $y = \beta_0 + \beta_1 x + \varepsilon$.

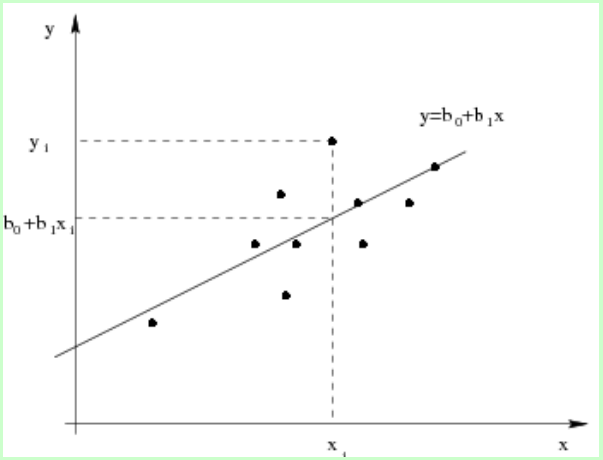
(Parametr β_0 interpretujeme jako teoretickou hodnotu Y při x = 0 a β_1 udává změnu Y, když X se změní o jednotku. Systém normálních rovnic získáme derivováním výrazu

$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ parciálně podle β_0 a β_1:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot 1 = 0, \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot x_i = 0$$

Řešením tohoto systému získáme odhady $b_0 = \dfrac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$, $b_1 = \dfrac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$

Po jednoduchých úpravách dospějeme ke tvaru $b_1 = \dfrac{s_{XY}}{s_X^2}$, kde s_{XY} je kovariance hodnot (x_i, y_i), $i = 1, ..., n$ a s_X^2 je rozptyl hodnot $x_1, ..., x_n$. Dále dostáváme $b_0 = \bar{y} - b_1 \bar{x}$, tedy regresní přímku můžeme vyjádřit ve tvaru $y = \bar{y} + \dfrac{s_{XY}}{s_X^2}(x - \bar{x})$.



Pro regresní přímku má reziduální součet čtverců tvar

$$S_E = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} x_i Y_i .$$

Odhad rozptylu: $S^2 = \frac{S_E}{\quad}$.

Index determinace:

$$ID = \frac{S}{\quad}, \text{ kde}$$

$$S_R = \sum \quad = \sum \quad + \quad \left(\quad \right)^2 = \frac{S_{XY}^2}{\quad} \sum \quad = \frac{S_{XY}^2}{S_X^2}, \; S_T = \sum \quad = \quad , \text{ tedy}$$

$$ID = \frac{n \frac{S_{XY}^2}{S_X^2}}{\quad} = \frac{S_{XY}^2}{\quad S_Y^2} = \quad .$$

Vidíme tedy, že v případě regresní přímky index determinace je roven kvadrátu koeficientu korelace.

Test významnosti směrnice regresní přímky (tj. test H_0: $\beta_1 = 0$ proti H_1: $\beta_1 \neq 0$) je ekvivalentní hypotéze o nulovosti koeficientu korelace (tj. testu H_0: $\rho = 0$ proti H_1: $\rho \neq 0$). Jestliže koeficient korelace veličin X, Y je blízký 0, nemá smysl počítat parametry regresní přímky.

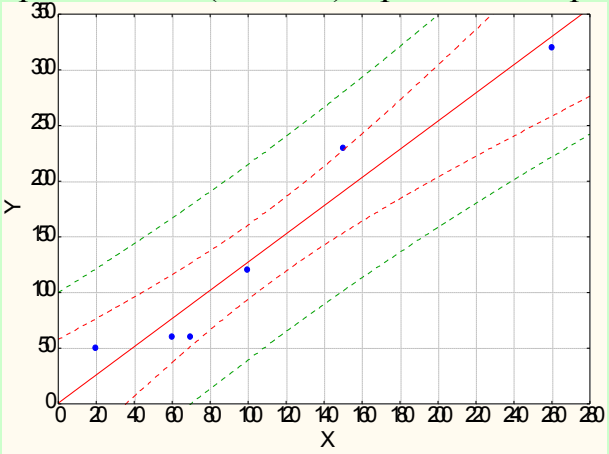
Interval spolehlivosti pro teoretickou regresní přímku při zadané hodnotě x_0 má meze:

$$d = \hat{a} + \hat{b}x_0 - t_{1-\alpha}(n-2)\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\,\hat{\sigma}, \quad h = \hat{a} + \hat{b}x_0 + t_{1-\alpha}(n-2)\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\,\hat{\sigma}.$$

Predikční interval spolehlivosti pro budoucí pozorování při zadané hodnotě x_0 má meze:

$$d = \hat{a} + \hat{b}x_0 - t_{1-\alpha}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\,\hat{\sigma}, \quad h = \hat{a} + \hat{b}x_0 + t_{1-\alpha}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\,\hat{\sigma}.$$

Data s proloženou regresní přímkou, pásy spolehlivosti (červeně) a predikčními pásy (zeleně)



Srovnání intervalu spolehlivosti a predikčního intervalu při zadané hodnotě x_0:

- oba intervaly jsou nejužší v místě $x_0 = m_X$,
- interval spolehlivosti pro dané x_0 je vždy užší než odpovídající predikční interval,
- predikční interval je určen pro individuální pozorování, zatímco interval spolehlivosti je určen pro hodnoty ležící na regresní přímce,
- s rostoucím rozsahem výběru se zmenšuje šířka obou intervalů.

Předpoklady použití regresní přímky:

- Závislost Y na X má lineární charakter.
- Pro celý rozsah uvažovaných hodnot nezávisle proměnné X je reziduální rozptyl s^2 konstantní (hovoříme o homoskedasticitě a znamená to, že variabilita hodnot závisle proměnné veličiny Y kolem regresní přímky je stejná pro všechny uvažované hodnoty nezávisle proměnné veličiny X).
- Hodnoty závisle proměnné veličiny Y mají normální rozložení pro dané hodnoty x_i a jsou stochasticky nezávislé (to souvisí s uspořádáním experimentu).

Poznámka: Menší odchylky od normality a homoskedasticity je možno tolerovat.

Sdružené regresní přímky

Uvažme nyní situaci, kdy obě veličiny Y a X jsou náhodné, přičemž samozřejmě předpokládáme, že X nezávisí na náhodné složce ε. Pak jde o případ oboustranné závislosti.

Závislost Y na X vystihuje 1. regresní přímka $Y = \beta_0 + \beta_1 X + \varepsilon$ a závislost X na Y vystihuje 2. regresní přímka $X = \gamma_0 + \gamma_1 Y + \varepsilon$. Odhady a_0, a_1 regresních koeficientů α, β v modelu $Y = \alpha + \beta x + \varepsilon$ získáme opět metodou nejmenších čtverců ve tvaru

$a_1 = \frac{s_{XY}}{s_X^2}, a_0 = \bar{y} - a_1 \bar{x} = \bar{y} - \frac{s_{XY}}{s_X^2} m_X$. 2. regresní přímka má tedy rovnici:

$$X - m_X = \frac{s_{XY}}{s_Y^2}(Y - m_Y).$$

1. a 2. regresní přímka se nazývají sdružené regresní přímky a odhady regresních koeficientů b_1, a_1 se nazývají odhady párově sdružených regresních koeficientů. Je zřejmé, že $b_1 a_1 = r_{XY}^2$. Rovnice sdružených regresních přímek můžeme tedy psát ve tvaru:

$$Y - m_Y = \frac{s_{XY}}{s_X^2}(x - m_X), \quad Y - m_X = r_{XY}^2 \frac{s_Y}{s_X}(x - m_Y).$$

Sdružené regresní přímky se protínají v bodě o souřadnicích $[m_X, m_Y]$. V případě, že náhodné veličiny X, Y jsou nekorelované, jsou odhady b_1, a_1 nulové a sdružené regresní přímky mají tvar $Y = m_Y$, $Y = m_X$. Pokud mezi náhodnými veličinami X, Y existuje úplná lineární závislost, pak sdružené regresní přímky splynou. K tomu dojde tehdy, když $r_{XY}^2 = 1$, tj. $a_1 = \frac{1}{b_1}$.

Označíme-li φ úhel, který svírají sdružené regresní přímky, pak z předešlých úvah plyne:

$\cos\varphi = 0$ <=> mezi X a Y neexistuje žádná lineární závislost;

$\cos\varphi = 1$ <=> mezi X a Y existuje úplná přímá lineární závislost;

$\cos\varphi = -1$ <=> mezi X a Y existuje úplná nepřímá lineární závislost.

Příklad:

Z fiktivního základního souboru všech vzorků oceli odpovídajících „všem myslitelným tavbám“ bylo do laboratoře dodáno 60 vzorků a zjištěny a hodnoty proměnné X – mez plasticity a Y – mez pevnosti. Datový soubor má tvar:

154	178	83	98	73	76
133	164	106	111	77	85
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

a) Určete regresní přímku meze pevnosti na mez plasticity.
b) Zakreslete regresní přímku do dvourozměrného tečkového diagramu spolu s 95% pásem spolehlivosti a predikčním pásem spolehlivosti.
c) Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
d) Vypočtěte index determinace a interpretujte ho.
e) Určete regresní přímku meze plasticity na mez pevnosti.
f) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
g) Obě regresní přímky zakreslete do téhož dvourozměrného tečkového diagramu.

Řešení v systému STATISTICA:

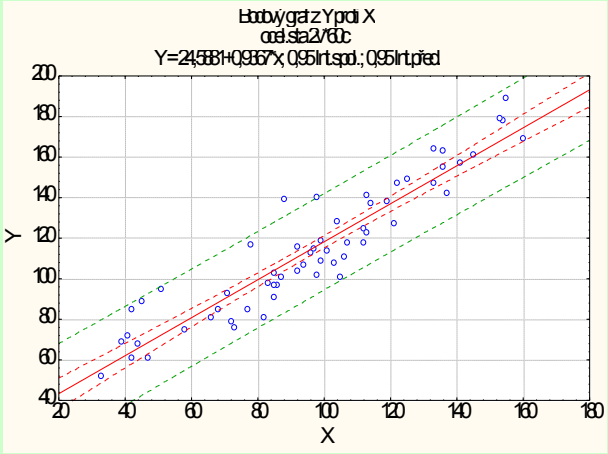
Odhad parametrů 1. regresní přímky:

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X - OK – OK – Výpočet: Výsledky regrese.

N=60	Výsledky regrese se závislou proměnnou R= ,93454811 R2= ,87338017 Upravené F(1,58)=400,06 p<0,0000 Směrod. chyba					
	Beta	Sm.chy beta	B	Sm.chy B	t(58)	Úroveň
Abs.čl			24,58	4,740	5,18	0,000
X	0,934	0,046	0,93	0,046	20,00	0,000

Zakreslení regresních pásů do dvourozměrného tečkového diagramu s proloženou regresní přímkou:

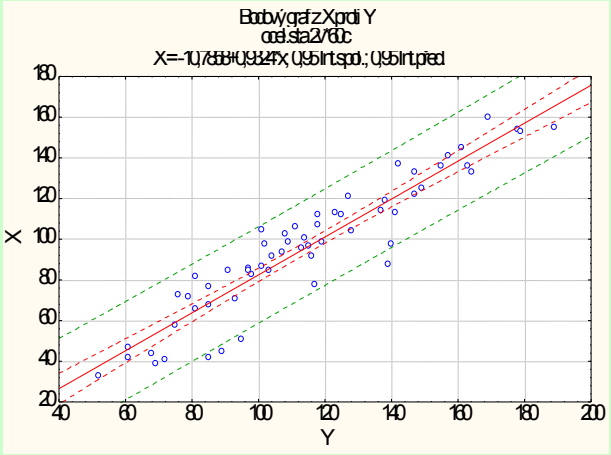
Grafy – Bodové grafy – Proměnné X, Y – OK – na záložce Detaily zaškrtneme Regresní pásy Spolehl. – OK. Ve vytvořeném grafu pak 2x klikneme na pozadí grafu a v nabídce Regresní pásy vybereme Přidat nový pár pásů – zaškrtneme Predikční.



Analogicky získáme výsledky pro 2. regresní přímku:

N=60	Beta	Sm.chyba beta	B	Sm.chyba B	t(58)	Úroveň p
Absc.den			-10,7858	5,54425	-1,9454	0,056575
Y	0,934548	0,046724	0,9324	0,046617	20,00161	0,000000

Výsledky regrese se závislou proměnnou: X (ocel.sta)
R= ,93454811 R2= ,87338017 Upravené R2= ,87119707
F(1,58)=400,06 p<0,0000 Směrod. chyba odhadu: 11,741



Nakreslení sdružených regresních přímek do jednoho diagramu:
K datovému souboru ocel.sta přidáme dvě nové proměnné y1 a y2. Do proměnné y1 uložíme predikované hodnoty meze pevnosti na mezi plasticity (do Dlouhého jména proměnné y1 napíšeme =24,58814 – 0,93668*x a do Dlouhého jména proměnné y2 napíšeme =(x+10,7858)/0,9324
Grafy – Bodové grafy – zaškrtneme Vícenásobný – Proměnné X: X, Y: Y, y1, y2 – OK. Ve vytvořeném grafu pak vypneme zobrazování značek pro y1, y2 a naopak zapneme Spojnici.